The NIST Speaker Recognition Evaluation

– Overview, Methodology, Systems, Results, Perspective –

George R. Doddington

Senior Principal Scientist

SRI International, Menlo Park, California, at

National Institute of Standards and Technology (NIST)

Gaithersburg, MD 20899, USA

Mark A. Przybocki and *Alvin F. Martin*

National Institute of Standards and Technology (NIST)

Gaithersburg, MD 20899, USA

Douglas A. Reynolds

MIT Lincoln Laboratory

244 Wood Street

Lexington, MA 02173, USA

Send correspondence to:

Alvin F. Martin

Technology Building 225 - Room A216

National Institute of Standards and Technology

Gaithersburg, MD 20899, USA

phone: (+1) 301 975-3169 – fax: (+1) 301 670-0939

alvin.martin@nist.gov – http://www.nist.gov/speech/martinal.htm

The NIST Speaker Recognition Evaluation

– Overview, Methodology, Systems, Results, Perspective –

57 pages

1 tables

13 figures

Abstract

This paper, based on three presentations made in 1998 at the RLA2C Workshop in Avignon, discusses the evaluation of speaker recognition systems from several perspectives. A general discussion of the speaker recognition task and the challenges and issues involved in its evaluation is offered. The NIST evaluations in this area, and specifically the 1998 evaluation, its objectives, protocols, and test data, are described. The algorithms used by the systems that were developed for this evaluation are summarized, compared, and contrasted. Overall performance results of this evaluation are presented by means of DET (Detection Error Trade-off) curves. These show the performance trade-off of missed detections and false alarms for each system, and the effects on performance of training condition, test segment duration, the speakers' sex, and the match or mismatch of training and test handsets. Several factors that were found to have an impact on performance, including pitch frequency, handset type, and noise, are discussed and DET curves showing their effects are presented. The paper concludes with some perspective on the history of this technology and where it may be going.

Keywords: Speaker recognition, identification, verification; Performance evaluation; NIST evaluations; DET (Detection Error Trade-off) curve

Résumé

Cet article, basé sur les trois exposés effectués en 1998 lors de la conférence RLA2C à Avignon, présente les méthodes de métrologie de la reconnaissance du locuteur. Apres un aperçu de ce qu'est la reconnaissance du locuteur et des problèmes posés, nous présenterons les objectifs, les méthodes, les données utilisées et les résultats de l'évaluation proposée par NIST en 1998. Pour cela nous utiliserons des DET-curves (Detection Error Trade-off). Ces courbes ont l'avantage de montrer le compromis entre erreur et fausse alarme et l'effet sur les performances des conditions d'entrainement, de la durée des segments de tests, du sexe du locuteur ou du type de combiné. Ensuite nous présenterons, comparerons et ferons la synthèse des differents algorithmes utilisés par les systèmes proposés par les participants.
Nous avons découvert que plusieurs facteurs influençaient fortement les performances des systèmes, comme par exemple la tonalité de la voix, le type de combiné utilisé ou le bruit ambiant. Nous les présenterons et mettrons en évidence à l'aide de DET-curves. Nous concluerons avec un historique des techniques de reconnaissance du locuteur et avec quelques projections de ce domaine dans l'avenir

Mots clé : Reconnaissance du locuteur, Identification, Vérification, Métrologie de performances, NIST, Courbes DET (Detection Error Trade-off).

1. Introduction

Language is the engine of civilization, and speech is its most powerful and natural form, despite the misappropriation of the term "*natural language processing*" to mean the processing of text. Textual language has certainly become extremely important in modern life, but speech has dimensions of richness that text cannot approximate. For example, the *health*, the *sex,* and the *attitude* of a person are all naturally and subliminally communicated by that person's speech. Such extra-linguistic information has social value and serves important communicative functions in our everyday lives.

Cues to a person's *identity* are also part of the extra-linguistic information communicated by the sound of a speaker's voice. And with the complexities of modern life, this ability to identify people by the sound of their voices finds value beyond the realm of personal living. With the advent of the information age and the high-powered but low-cost computers that fuel it, speaker recognition has become an attractive potential capability – albeit one that largely remains to be fulfilled through research and technology development efforts.

1.1. NIST Evaluations

The National Institute of Standards and Technology (NIST), located in Gaithersburg, Maryland, USA, has played a vital role in working with industry to develop and apply technology, measurements, and standards, for several years. The Spoken Natural Language Processing group of NIST's Information Technology Laboratory conducts annual benchmark tests to evaluate the state-of-the-art in the areas surrounding core speech recognition technologies.

NIST has coordinated an ongoing series of Speaker Recognition evaluations, which have provided an important contribution to the direction of research efforts and the calibration of technical capabilities. They are intended to be of interest to all researchers working on the general problem of text independent speaker recognition. To this end, the evaluations are designed to be simple, to focus on core technology issues, to be fully supported, and to be accessible. A follow-up workshop for the participants to discuss research findings is held after each evaluation. Participation in these evaluations is solicited for all research sites that find the task and the evaluations of interest.

1.2. Outline

The authors have been involved in the NIST speaker recognition evaluations as planners, organizers, scorers, sponsors, and participants. Section 2 of this paper (primary author G. Doddington) offers a somewhat personal overview of evaluation methodology for speaker recognition based on long experience in the field. Section 3 of this paper (primary authors M. Przybocki and A. Martin) discusses the objectives and protocols of this evaluation. Section 4 (primary author D. Reynolds) describes the various systems from laboratories in the United States and Europe that were included in the evaluation. Section 5 (primary authors M. Przybocki and A. Martin) presents a number of charts of the performance results of the evaluation, with some contrast with the results from the evaluation of the previous year. We discuss two specific factors that we found to significantly affect the level of performance, namely speakers' average pitch and the handset types of the phones used. Section 6 summarizes the NIST evaluation process and suggests future directions for research for where this technology is headed.

2. Overview of Evaluation Methodology

The increase in application opportunities has resulted in a heightened interest in speaker recognition research. A key element in this research is the identification of key research tasks and the establishment of evaluation methodology to support them. The purpose of this section is to provide an interpretive overview on speaker recognition tasks and evaluation methodology as previously presented at the "La Reconnaissance du Locuteur et ses Applications Commerciales et Criminalistiques" (RLA2C) Workshop [4].

2.1. Applications

There are numerous possible ways of categorizing different applications of speaker recognition. One meaningful way is to divide the applications into two groups according to the immediate beneficiary of the process. Specifically, the question is whether it is the *speaker* who benefits, or someone *else*. This split is particularly useful because it also has a profound impact on the task definition, the technology, the system design and system performance.

Examples of speaker recognition applications that serve the speaker are security systems that provide control of physical entry and information access. Such systems are in general more intricate in design and capable of better performance than other kinds of systems. For example, because security system users are cooperative, they can collaborate with the system by saying what is asked of them and by proffering their identities to make the recognition task easier. Also, special care is taken in the design of these systems to ensure good speaker recognition performance – for example, by using high quality uniform microphones and by controlling the ambient noise level, where possible.

Examples of speaker recognition applications that serve someone other than the speaker include forensic applications. And while use of speaker recognition in such cases is less encumbered by design details, performance is usually significantly worse – the speaker may be using an unknown microphone (often a carbon button telephone handset) or may be particularly emotional, and the recording is typically of low quality for a variety of reasons. The one advantage that forensic type applications often do have is in the amount of speech data. Whereas in security applications, where time is of the essence and speech segments are measured in a few seconds, forensic applications may have minutes of speech to use.

2.2. Task Definitions

In order to conduct a productive applied research effort directed toward speaker recognition technology, it is helpful to have a clear understanding and expression of the research objective. This research objective is properly expressed in terms of a formal definition of the speaker recognition task.

Ideally, a formal task definition of speaker recognition will serve more than to organize research within a single research effort. It can also serve to share resources and results across a number of different research sites. This tendency toward globalization of speech research has accelerated strongly during the past ten years, due in part to a gradual realization of the immense challenge of understanding speech and creating useful speech technology.

There are logically two parts to defining the speaker recognition task. These are to define the nature of the speaker recognition decisions, and to define the nature of the speech data on which these decisions are to be made.

2.2.1. Recognition Decisions

Speaker recognition is the general term used to include all of the many different tasks of discriminating people based on the sound of their voices. There are many terms used to distinguish different tasks, including speaker identification, speaker verification, speaker spotting and speaker detection. But it will be most useful to focus primary discussion on just two different kinds of tasks, identification and verification.

2.2.1.1. Identification

Speaker identification is the task of deciding, given a sample of speech, who among many candidate speakers said it. This is an N-class decision task, where N is the number of candidate speakers. And of course N has a powerful influence on the ultimate performance for an identification task – performance might be extremely good for just a few speakers (especially if they happen to have distinctly different voices), but as N becomes arbitrarily large, the probability of correctly identifying the speaker becomes arbitrarily small and approaches zero.

There are variants of the identification task, including most notably the situation where the actual speaker may be *none* of the candidate speakers. The identification task is called a "*closed set*" task when the actual speaker is always one of the candidate speakers. When the actual speaker may not be one of the candidate speakers, it is called an "*open set*" task.

In most cases, the speaker identification task appears to be more of a scientific game than an operationally useful capability. This is especially true for security system type applications, because here the speaker has good reason to cooperate with the system by proffering an identity, thus reducing the N-class task down to a 2-class task. Even for forensic type applications, the problem most often is one of evaluating each candidate separately rather than choosing one from among many candidates. Thus it seems that, although the speaker identification task may garner considerable scientific interest, it is the speaker verification task that has the greatest application potential.

2.2.1.2. Verification

Speaker verification is the task of deciding, given a sample of speech, whether a specified candidate speaker said it. This is a 2-class task and is sometimes referred to as a speaker detection task. The two

classes for the verification task are: the specified speaker (known variously as the "*true speaker*" or the "*target speaker*"); and some speaker other than the specified speaker (known variously as an "*impostor*" or a "*non-target speaker*").

One very nice characteristic of the speaker verification task, deriving from the inherent nature of a 2-class task, is that the performance is independent of the number of potential impostor speakers. Thus the measured performance of a verification system is independent of the number of impostor speakers in the test set. (This assumes that the verification system doesn't "know" the impostor speakers. Improved impostor resistance can be achieved for small numbers of impostors if the system has explicit knowledge of the impostors' speech.) This does not imply, however, that performance is independent of the *choice* of impostors – if impostors similar to the target speaker are selected, then naturally performance will be worse.

2.2.2. Operating Modes

Generally speaking, there are two different modes of speech used in speaker recognition, which correspond to the two different kinds of applications. When the speaker is cooperative, the system may know what the speaker is supposed to say, and better performance may be achieved. When the speaker is uncooperative or unaware, then the system will be handicapped by lack of this knowledge.

2.2.2.1. Text-dependent recognition

A speaker recognition system is called "*text dependent*" if it knows what the speaker is supposed to say. In the typical text dependent recognition scenario, the speaker will either say a predefined utterance, or will be prompted to say an utterance by the system. In either case, the target speaker will have explicitly spoken these words during an enrollment session.

This explicit knowledge can be used to good effect in building detailed dynamic models of the speaker. These models can be word- and phone-specific and thus can calibrate the target speaker with great accuracy – sometimes too accurately, if the speaker doesn't produce the appropriate utterance properly! But, with nominal speaker cooperation, text-dependent recognition improves recognition performance while at the same time minimizing the amount of speech data required.

2.2.2.2. Text-independent recognition

A speaker recognition system is called "*text independent*" if it doesn't have foreknowledge of what the speaker is saying. This forces the technology to deal with whatever speech data happens to be available, both for training and for recognition decisions. Because of this, text independent systems typically exhibit worse performance than do text dependent systems, at least per unit of speech.

There is a significant research advantage to working on text independent recognition, beyond just the ability to use significantly larger amounts of speech data. This is namely the general absence of idiosyncrasies in the definition of the task and the specification of the speech data. This allows simple and general task and data specification, which in turn supports much broader collaboration and sharing of results than has been realized in text-dependent recognition. Partially as a result of this advantage, interest in text independent recognition has risen significantly.

2.3. The Technical Challenges

People's voices are distinctive. That is, a person's speech exhibits distinctive characteristics that indicate the identity of the speaker. We are all familiar with this, and we all use it in our everyday lives to help us interact with others. Of course, from time to time we might notice that one person sounds very much like another person we know. Or we might even momentarily mistake one person as another because of the sound of the person's voice. But this similarity between voices of different individuals is not what the technical challenge in speaker recognition is all about.

The challenge in speaker recognition is *variance*, not similarity. That is, the challenge is to decode a highly variable speech signal into the characteristics that indicate the speaker's identity. These variations are formidable and myriad. And the principal source of variance is the speaker.

2.3.1. The Speaker

While it is the speaker's voice that provides the recognition capability, it is also the speaker's variability that makes the problem so difficult. An explanation for why the speaker's variability is such a vexing problem is that the use of speech, unlike fingerprints or handprints or retinal patterns, is to a very large

degree a result of what the person *does*, rather than who the person *is* – speech is a performing art, and each performance is unique.

Some of the sources of variability are within the speaker's control. Some are not. Here is a list of some of the factors that have been offered as explanations for speaker variability leading to less than perfect speaker recognition performance:

- The session – Early experiments showed good speaker recognition performance when training and testing were both conducted in the same session. The single session experiment successfully avoids these problems, but unfortunately does not contribute to solving them.
- Health – Respiratory infections are a definite problem, and of course laryngitis is the ultimate health problem for speaker recognition. Other factors that might be considered part of a person's health include emotional state and metabolic rate.
- Educational level and intelligence – Although this subject is taboo, general intellectual keenness may play a role, at least in cooperative systems where speaker control can have a beneficial effect on performance.
- Speech effort level and speaking rate – These factors are at least superficially controllable, although people are not typically conscious of them. This is especially true for the Lombard effect (where people unconsciously talk louder when exposed to auditory noise). Such changes in speech production cause complex changes in the speech signal beyond simple energy and rate changes.
- Experience – This applies only to cooperative systems where the speaker has the opportunity to interact with the speaker recognition system repeatedly. In these cases, it has been observed that there is a striking learning effect. How this learning effect is divided between the machine learning the speaker and the speaker learning to talk to the machine is not totally clear, however.

2.3.2. Other Challenges

There are other factors, beyond speaker variability, that present a challenge to speaker recognition technology. These deal with problems in getting the acoustical signal intact from the speaker to the recognition system:

- Acoustical noise – This may be a large or small problem, depending on the microphone used to transduce the speech signal and the acoustical environments within which the system is required to perform.
- The microphone – Second only to the speaker, the microphone and associated sound pick-up issues probably represent the next biggest challenge to speaker recognition. This is especially true for applications that use the telephone, because of the wide variety of telephone handsets and microphone elements. People also have a wide range of ways that they hold telephone handsets and may continually change positions while talking, presenting problems to human interlocutors as well as speaker recognition systems.
- Electrical transmission – This has been a problem in the past, and current mobile telephone systems continue to pose a challenge. However, improvement in line quality and the rapid deployment of high performance digital communication systems may largely eliminate transmission from the palette of problems that face speaker recognition in the future.

2.4. Performance Measures

Performance measures serve a number of purposes. These include, most importantly, a means for evaluating research ideas and making consistent long-term technical progress. Other reasons include comparing different systems, evaluating the effectiveness of technology for specific applications, marketing research to sponsors and selling products to customers.

Performance measures need to be easy to understand and clear. In deciding how to express performance figures, there is good reason to choose error over accuracy. The reason is that error lends itself to better intuition than does accuracy. For example, it is easy to see that reducing the error rate from 10 percent to 5 percent is in a very real sense a doubling of performance, which is of course the same as increasing the accuracy from 90 percent to 95 percent.

2.4.1. Identification

The speaker identification task is straightforward and lends itself to a simple bottom-line identification error rate, E_{ID}, as the performance measure:

$$E_{ID} = n_{err} / n_{tot},$$

where n_{tot} and n_{err} are the total number of trials and the number of trials in error, respectively. The error rate may vary with speaker and other conditions, but the basic performance measure is quite straightforward.

2.4.2. Verification

Speaker verification is a detection task, and therefore there exists considerable support for it, in terms of existing conventions and procedures for measuring the performance of detection systems.

2.4.2.1. Miss/False Alarm

Detection system performance is usually characterized in terms of two error measures, namely the miss and false alarm error rates. These correspond respectively to the probability of not detecting the target speaker when present, E_{miss}, and the probability of falsely detecting the target speaker when not present, E_{fa}. These measures are calculated as:

$$E_{miss} = n_{miss} / n_{target},$$

where n_{target} and n_{miss} are the number of target trials and the number of those where the target speaker was not detected, respectively, and

$$E_{fa} = n_{fa} / n_{impostor},$$

where $n_{impostor}$ and n_{fa} are the number of impostor trials and the number of those where the target speaker was falsely detected, respectively.

2.4.2.2. Equal Error Rate

Miss and false alarm rates, while properly characterizing the performance, still do not produce a single number performance figure. The use of equal error rate essentially combines miss and false alarm rates into a single number by finding the decision threshold at which they are equal. This only works if the decision threshold is adjustable, of course.

2.4.2.3. Geometric Mean Error

While equal error rate may be a reasonable performance measure for laboratory systems, this doesn't extend so easily to fielded systems. Operational systems do not exhibit equal miss and false alarm error rates, and usually these are purposely adjusted to be quite different from each other. Nonetheless, it would be desirable to compare different systems in terms of a single error figure that represents their underlying ability to discriminate among speakers.

The geometric mean error, E_{GM}, is a statistic that offers the desired ability to compare systems at other than the equal error point. It is defined simply as

$$E_{GM} \equiv \sqrt{E_{Miss} \cdot E_{False\ Alarm}}\ .$$

E_{GM} will be fairly constant as miss and false alarm rates vary, at least over a modest range of values. And while it may be used as a rough comparison of different systems, probably a better use for E_{GM} is as a diagnostic measure of performance during research or system upgrade. By this means, for example, small variations in miss/false alarm trade-off need not prevent a meaningful comparison.

E_{GM} is not perfectly constant, however. Typically, it tends to increase as the miss rate is decreased. This effect usually becomes more pronounced when the miss rate gets down to about one or two percent. This gives rise to the adage that "*it's easier to reject impostors than it is to accept true speakers*". This droll folk wisdom reflects the behavioral statistics (e.g., state of health and emotions) of the true speakers – an interesting statistical commentary on the inherent variability of the human voice.

2.4.2.4. Detection Cost

Another good approach to representing system performance as a single number is to formulate a detection cost function. This is a weighted arithmetic mean of the miss and false alarm rates. This measure has the advantage that it models the application and produces a number, which is directly meaningful to the application. Detection cost, C_{det}, is typically modeled as a weighted sum of the posterior probabilities of miss and false alarm:

$$C_{\det} = c_{miss} \cdot E_{miss} \cdot P_{target} + c_{fa} \cdot E_{fa} \cdot \left(1 - P_{target}\right),$$

where c_{miss} and c_{fa} are the costs of a miss and a false alarm, respectively, and P_{target} is the a priori probability of a target.

The detection cost function as defined above seems to be an excellent evaluation measure, because it quantifies the value (cost) of the technology to the application. And for many different applications, the speaker recognition value/cost is arguably well represented by this detection cost function. Yet there appears to be little interest in the scientific community for such abstract measures. Researchers prefer to understand miss and false alarm error trade off.

2.4.3. The DET Plot

The trade-off between miss and false alarm error rates have traditionally been shown on so-called ROC curves, which plot detection probability as a function of false alarm probability. An improvement to this visualization aid, called the DET plot (for Detection Error Trade-off), has been introduced by NIST recently [15]. The DET plot improves the visual presentation of detection error trade-off by plotting miss and false alarm probabilities according to their corresponding Gaussian deviates, rather than the probabilities themselves. This results in a nonlinear probability scale, but the advantage is that the plots are visually more intuitive. In particular, if the distributions of error probabilities are Gaussian, then the resulting trade-off curves are straight lines. And the distance between curves depicts performance differences more meaningfully. Examples of DET plots may be seen below in section 5. In particular, Figure 3 contrasts a DET plot and a traditional ROC plot for the same data.

2.4.4. Decision Making

Speaker recognition technology deals with making decisions. The task is quite simple – to make a decision about the identity of the speaker, based on the sound of the person's voice. That seems clear enough. Yet the actual decision making process, specifically the setting of decision thresholds, has often been neglected in speaker recognition research.

Making these decisions has often been dismissed as an unchallenging problem to be addressed during application development. Yet for those who actually have had to deploy real operational systems, the problem has been found to be quite challenging, indeed. For the case of cooperative speaker systems,

performance is invariably found to vary widely among different speakers. But homogeneous performance, with nice low miss rates for everyone, is a system requirement. So researchers endeavor to calibrate each speaker separately, only to find that what is observed in training may be destructively misleading in usage – speaker normalization sometimes results in *worse* performance!

The important point here is that the actual decision process must be considered to be part of any comprehensive speaker recognition research program and that the ability of a system to make good decisions should be an integral part of the evaluation.

2.4.5. Pooling Data

The issue of how to pool data is related to the problem of setting thresholds and making decisions. Pooling data for different target speakers can be especially painful, because the good performance given by speaker-dependent thresholds may dissolve into poor performance from a muddled threshold after results have been pooled.

One suggested solution to this problem is to find the equal error rate for each target speaker and then to average these error rates over all speakers. While this procedure invariably shows superior performance, the result is illusory. There are several reasons why this is a bad practice. The first is that knowledge of the best decision thresholds is inadmissible information – decisions must be made without the artificial luxury of knowing the answers. The second is that inference of the best decision thresholds from the test data is false knowledge – random sampling produces the illusion of good decision points that do not really exist. This can have a strong bias on results whenever there is a relatively small amount of data for each of the target speakers, which is usually the case.

2.5. Evaluation

Having a measure of performance is certainly desirable, even necessary, in order to evaluate a system. But performance is a function of many factors. Consideration of these factors, and how they influence performance, is necessary in order to understand the technology and apply it successfully.

2.5.1. Evaluation Factors

Evaluation factors are factors that influence the performance of the system. Ideally we would like to vary each of these factors and evaluate their effect so as to understand how they relate to performance. Here are some of the most important evaluation factors:

- Amount of target training data – The more training data that a system has in order to learn a speaker's voice characteristics, the better will be the performance of the system. In particular, data from a number of different sessions is desirable, because a speaker's voice characteristics change significantly from session to session. But for cooperative speaker systems there is usually an overriding need to minimize the demands on the speaker and thus the amount of training data. And while multi-session training would be very helpful, this is definitely a disadvantage to the application.

- Test segment duration – This is probably the most studied factor in speaker recognition performance, with longer segment durations providing significantly better performance. This is discussed below (section 5.2) for the 1998 NIST evaluation.

- Microphone differences – Microphone differences are one of the most serious problems facing speaker recognition, especially when dealing with the telephone, where Edison's old nonlinear carbon-button microphone still represents a significant fraction of all transducers. The importance of microphone differences may be illustrated by comparing the performance obtained when training and testing are performed on a single type of microphone with that obtained when training and testing are performed on different types of microphones. This may be done easily enough when using telephone data, because there are two different types of microphone elements that are commonly used in telephone handsets. These are namely the traditional carbon-button element, and the now more common electret element, which is rapidly replacing carbon buttons because of its lower cost. This contrast is discussed below (section 5.3) for the 1998 NIST evaluation.

- Noise –Noise and distortion will obviously corrupt performance, but there has been little systematic study of this issue. See section 5.5 below for a discussion of NIST's effort to correlate subjective noise with system performance in the 1998 evaluation.

- Temporal drift – Temporal factors play an important role in speaker recognition performance. Depending on the needs of the intended application, evaluation may need to stretch out over months or even years. For systems that include adaptation to speaker changes, this may not be a critical factor. However, it may also be that different age groups exhibit different performance characteristics. For cooperative speaker systems, special attention needs to be paid to the period immediately following a speaker's enrollment. This is an especially difficult time, because the system must bridge the gap between an unadapted and relatively inaccurate model and well-adapted model of the speaker at the same time that the speaker is learning how to speak to the system.
- Sex differences – Male and female voices are generally quite different from each other, both in physical characteristics (pitch and vocal tract length, for example) and in linguistic and stylistic use. For this reason, it is usually informative to measure performance separately for men and women. This is discussed below (section 5.4) for the 1998 NIST evaluation.

2.5.2. Evaluation Methodology

The two different types of speaker recognition applications, namely cooperative (text dependent) recognition and tacit (text independent) recognition, have a profound influence on the design of an evaluation. For the text dependent case, the task is invariably system-specific and usually very idiosyncratic, making system evaluation specific to the system. This in turn inhibits the comparison of performance among different systems.

2.5.2.1. Statistical Significance

The overarching consideration in designing an evaluation is statistical significance. An evaluation is pointless if it is not significant. Furthermore, if the evaluation is to be helpful to research, then statistical significance must be with respect to multiple and various selected conditions. For example, if only one target speaker is used, then statistically significant results may be obtained, but unfortunately they will be valid only for that particular speaker. Clearly, if statistical significance across a population in general is required, then an adequate number of speakers must be sampled. Regardless of the number of speakers sampled, however, care must be exercised to ensure that the sample population represents the population of interest. For example, it might be particularly easy to recruit college students, and so an experiment might

be limited to college students. Are the results of this experiment valid for populations of different age and educational demographics? Perhaps. Perhaps not. Assertions of statistical significance are dangerous in such circumstances.

One of the difficulties facing evaluation is the large number of factors that influence performance and the complex way that they may interact. Requiring meaningful (i.e., statistically significant) results for many different combinations of evaluation conditions easily results in an unmanageably large test.

2.5.2.2. The Rule of 30

In determining the required size of a corpus, a helpful rule is what might be called "*the rule of 30*". This comes directly from the binomial distribution, assuming independent trials. Here is the rule:

> *To be 90 percent confident that the true error rate is within +/- 30 percent of the observed error rate, there must be at least 30 errors.*

This confidence interval and proportional bound on error rate are reasonable values that yield reasonable requirements for the size of an evaluation corpus. The rule may be applied by considering the performance goals or expectations for the evaluation. For example, suppose that the performance goals are 1 percent miss and 0.1 percent false alarm. Thirty errors at 1 percent miss implies a total of 3,000 true speaker trials, and thirty errors at 0.1 percent false alarm implies a total of 30,000 impostor trials. Note, however, that a key assumption in these calculations is that the trials are *independent*. The implications of this can be daunting – does this mean 3,000 true speakers and 30,000 impostors? Strictly speaking, it probably does. The alternative is to compromise on independence and not to take assertions of statistical significance too seriously.

2.5.2.3. Speaker Selection

Target speakers should be chosen so as to best represent the intended application. This includes selection of sex, age, dialect and other demographic factors of potential relevance. And of course, there must be enough of them. This is especially important because of performance inhomogeneities among speakers. Specifically, there will be differences in performance among speakers for which there are no clear explanations or underlying mechanisms. This has led to the jocular characterization of the target

population as being composed of "*sheep*" and "*goats*". In this characterization, the sheep are well behaved and dominate the population, whereas the goats, though in a minority, tend to determine the performance of the system through their disproportionate contribution of errors.

The impostor population should model the target population, which is typically done in faultless style by having the targets serve also as impostors. Like targets, impostors also have barnyard appellations, which follow from inhomogeneities in impostor performance across the speaker population. Specifically, there are some impostors who have unusually good success at impersonating many different target speakers. These are called "*wolves*". And then there are some target speakers who seem unusually susceptible to many different impostors. These are called "*lambs*". The purpose of these appellations is to point to the need to study and understand these speaker inhomogeneities. For an initial effort in this direction, see [5].

Another profitable direction for speaker selection would be so as to focus research on idiosyncratic voice differences. Without special selection, many if not most of the salient differences in voices will be broad differences such as age, size and dialect. In order to focus research effort on fine idiosyncratic voice differences, it would be helpful to select a population of speakers who all share the same gross physical and linguistic characteristics.

2.5.2.4. Data Collection

For target speakers, test data should be collected in many different sessions. Performing multiple tests from data collected in a single session is of limited value at best, because of the strong correlation of speech characteristics within a single session. Test data should also be processed chronologically, so that test data always follows training data.

Impostor trials are easier to come by if it is possible to use a single impostor test segment to test against multiple target speakers. Because of this multiplicative effect, it may be possible to run an extremely large number of impostor trials. The impulse to do this should be resisted, because the statistical significance of the results is determined in a fundamental way by the number of target speakers, not simply by the number of trials performed on them. So running an excessive number of impostor trials is just wasteful.

By convention and common sense, cross-sex impostor trials are to be avoided. Although allowing cross-sex trials will improve the apparent performance of a system somewhat, it confounds comparison of performance with other systems, and it risks straining the credulity of sponsors, colleagues and bystanders.

3. The 1998 Evaluation: Objectives, Data, Participants

This section summarizes the protocols for the 1998 NIST evaluation [18] and offers some contrasts with those for the 1997 evaluation [17]. We discuss the overall technical objective, the evaluation metric and how results were presented, the evaluation data set, the training and test conditions evaluated, and the participating sites. The official plans for these evaluations are posted on the NIST Spoken Natural Language Processing group's website [17, 18].

3.1. Technical Objective

The NIST evaluations have focused on the task of speaker detection (or equivalently, speaker verification). That is, the task has been to determine whether a specified target speaker is speaking during a given speech segment.

This task has been posed in the context of conversational telephone speech and for limited training data. The evaluations are designed to foster research progress, with the goals of:

- Exploring promising new ideas in speaker recognition,
- Developing advanced technology incorporating these ideas, and
- Measuring the performance of this technology.

Speaker detection performance has been evaluated by measuring the correctness of detection decisions for an ensemble of speech segments. These were segments selected to represent a statistical sampling of the conditions of evaluation interest. For each of these segments a set of target speaker identities was assigned as test hypotheses. Every hypothesis was to be judged as either true or false. Each decision had to be based only upon the specified test segment and target speaker. Use of information about other test segments and/or other target speakers **has not** been allowed.

3.2. Evaluation Measure

The formal evaluation measure was the detection cost function (see 2.4.2.4), defined as a weighted sum of the miss and false alarm error probabilities:

$$C_{\text{det}} = c_{miss} \cdot E_{miss} \cdot P_{target} + c_{fa} \cdot E_{fa} \cdot \left(1 - P_{target}\right),$$

The parameters of this cost function are the relative costs of detection errors, c_{miss} and c_{fa}, and the a priori probability of the target, P_{target}. The evaluations have used the following parameter values:

$$c_{miss} = 10;\ c_{fa} = 1;\ P_{target} = 0.01$$

In addition to the (binary) detection decision, a decision score was also required for each test hypothesis. The decision scores were used to produce Detection Error Trade-off (DET) curves (see 2.4.3), showing the trade-off of misses and false alarms.

3.3. Evaluation Data

The speech data came from Phase 1 of the Switchboard-2 corpus in 1997, and from Phase 2 in 1998. Switchboard-2 is a corpus of conversational telephone speech that was created using a collection paradigm similar to that of the original Switchboard corpus. These corpora are available from the Linguistic Data Consortium (LDC) [14].

As with Switchboard, Switchboard-2 speakers were assigned topics to discuss with another speaker, whom they did not know. In contrast to the earlier corpus, however, they were given permission to ignore the assigned topic, and in most cases they chose to do so. These speakers were generally college students, considerably younger on average than the Switchboard speakers, and much of the dialogue is dominated by what may be characterized as “college chit-chat”. The Phase 1 speakers were primarily from the northeastern United States, while the Phase 2 speakers were primarily from the American Midwest.

The evaluation data contained about 400 target speakers in 1997, and about 500 in 1998. There were about 5000 test segments per duration both years. Ten or eleven targets were specified for each segment to be tested against.

New in 1998 was the creation process of the test and training segments. An automatic speech detector determined where the speech signal was present in a conversation. NIST manually audited these segments in 1997, rejecting ones that did not meet the acceptance criteria (rules designed to eliminate non-speech segments). Then the appropriate amount of speech was concatenated together, eliminating silences. It was determined in a special test after the 1997 evaluation that removing the auditing for non-speech segments had little effect on recognition performance. For 1998, therefore, the segment selection process was essentially entirely automatic, with human verification only that the speakers were correct. Segments that were noisy or lacking in speech (see 5.5 below) were noted for possible later analysis, but were not removed.

The development data set for the 1997 evaluation was the data used in the 1996 evaluation, while for the 1998 evaluation the 1997 data served this role. These three data sets, which may be used as kits for research efforts, are now available from the LDC.

3.4. Evaluation Conditions

The 1998 evaluation officially consisted of nine different tests. Participating sites could choose to do some or all of these, but were required to process all segments and targets specified in each test. The nine tests were defined by three different training conditions, and by test segments of three different durations. Each test included separately male speakers and test segments and female speakers and test segments. There were no cross gender tests.

3.4.1. Training

There were 3 training conditions for each target speaker in the 1998 evaluation. The 3 conditions were:

- "One-session" training – The training data consists of 2 minutes of speech data taken from only one conversation.
- "Two-session" training – The training data consists of 1 minute of speech data taken from each of two different conversations collected using the same telephone number (and presumably the same telephone handset).

- "Two-session-full" training – The training data consists of all speech data available in the two conversations used for two-session training.

The 1997 evaluation, it may be noted, used the first two of these training conditions. It did not use the two-session-full training condition, but it also included the following condition:

- "Two-handset" training – The training data consists of 1 minute of speech data taken from each of two different conversations collected using different telephone numbers (and thus presumably different telephone handsets).

Table 1 includes a summary of the training conditions.

3.4.2. Test

Performance was computed and evaluated separately for female and male target speakers and for the 3 training conditions. The test segments included in the evaluation were organized by sex and by duration. The evaluation plan specified three factors of interest with respect to which performance would be examined. They were:

- Test segment duration – Performance was computed separately for the 3, 10, and 30 second test segments.
- Same/different number – Performance was computed separately for test segments that used the same phone number as was used for training the true speaker versus those segments which used a different phone number.
- Same/different handset type - Performance was computed separately for different number test segments with the same handset microphone type label (either electret or carbon-button, as discussed below) as the true speaker training data versus those segments with a different handset microphone type label

The test segments came from different conversation sides from the training data. Only one segment of each duration was selected from a conversation side. The 10-second segment was a subsegment of the 30, and the 3-second segment was a subsegment of the 10. For convenience, however, they were provided as separate files in the test data.

The 1997 evaluation had focused on different number tests, and the same/different type of each test segment was kept unknown to the system. The 1998 evaluation sought to de-emphasize handset differences and thus focused on same number tests. The system was informed of whether each test segment involved a same or different number test with respect to the true speaker. Note that all tests involving a speaker other than the true speaker of a test segment (non-target tests) necessarily involved different handsets in training and test.

Also in 1998, and in contrast to 1997, systems were told the classification of all training and test handsets as being of either carbon button or electret microphone type as determined by the MIT Lincoln Lab handset classifier. The importance of handset type to recognition performance had been observed in the previous evaluation (section 5.3).

Table 1 includes a summary of the test conditions.

3.5. Participants

Twelve research sites participated in the 1998 evaluation. Five of the European sites, while submitting results for separate systems, worked cooperatively in what was called the ELISA Consortium. The types of systems developed by many of these sites are discussed in section 4. The performance results presented in section 5 do not, however, identify the individual sites. The sites and their designations (* indicates ELISA Consortium participant) were:

- A2RT – Department of Language and Speech - Nijmegen University - Netherlands
- BBN – BBN Technologies - GTE - Cambridge, MA - USA
- CIRC* – Circuits and Systems Group - Ecole Polytechnique Federale de Lausanne (EPFL) - Switzerland
- Dragon – Dragon Systems, Inc. - Newton, MA - USA
- ENST* – Ecole Nationale Superieure des Telecommunications - Paris - France
- IDIAP* - Institut Dalle Molle d'Intelligence Artificielle Perceptive - Martigny - Switzerland
- IRISA* – Institut de Recherche en Informatique et Systemes Aleatoires - Rennes - France

- LIA* – Laboratoire Informatique - Universite d'Avignon et des Pays de Vaucluse - France
- LIP6 – Laboratoire d'Informatique de Paris 6 - Universite Pierre et Marie Curie - France
- MIT-LL – MIT Lincoln Laboratory - Lexington, MA - USA
- OGI – Oregon Graduate Institute of Science and Technology - Portland, OR - USA
- SRI – Speech Technology and Research Laboratory - SRI International - Menlo Park, CA - USA

While not a participant in the 1998 evaluation, Ensigma Ltd., located in Chepstow, UK, was represented at the review workshop following the evaluation since it had participated in previous evaluations and presented some performance results on its then current speaker recognition system. Its system is discussed along with those of the evaluation participants in section 4.

4. The 1998 Evaluation: Technology Overview

In this section we provide a high-level overview of the recognition technology employed by participants in the 1998 NIST evaluation. The aim is to outline the general technology trends of state-of-the-art, text-independent, speaker recognition systems in the areas of features, speaker models and score normalization. We also discuss the use of fusion systems and potential future research directions. It is outside the scope of this article to describe the different approaches in detail, for which the reader is directed to the referenced papers or cited evaluation participant. In addition, association of a site with a particular technique does not imply that they invented the approach, merely that this is what they used in their eval98 system.

While based on a presentation at the RLA2C Workshop, the material in this section has not been published previously.

4.1. Canonical Recognizer Structure

From a general perspective, the approach used by all systems in the NIST speaker recognition evaluation is that of a likelihood ratio detector. The canonical structure of this general system is shown in Figure 1 and consists of three main components. The first is the front-end processing which represents the signal processing to extract features used for model training and recognition and any signal related channel compensation to minimize the effects of different channels on the features. The second component is the

target speaker and background models used in forming the likelihood ratio statistic for a test utterance. The final component is a post-processing step of score normalization usually applied to stabilize detection scores so that speaker-independent thresholds are more effective. Although described as distinct components, in some systems these components can be integrated thus blurring the category to which they belong. For example, channel compensation techniques can and do occur in front-end processing, in model training and in score normalization. We next describe the trends in these different areas.

4.2. Speech Features

Of the major system components presented in Figure 1, the greatest commonality of approach among the participating sites was seen in the front-end processing. A majority of systems employed the following standard set of front-end processing steps

- **Signal bandlimiting** – Systems used only spectral information from the frequency range 300-3400 Hz, the voice band of the telephone signal.
- **Cepstral feature extraction** – Most systems used filterbank magnitude spectral representations followed by transformation to the cepstral domain using the DCT (discrete cosine transform). A few used LPC (linear predictive coefficients) spectral representations followed by recursive equation expansion of cepstral coefficients.
- **Cepstral derivatives** – Generally first order delta cepstral features were appended to the static cepstral feature vector. One system used delta cepstra as an independent feature stream. Second order derivatives were not found to provide measurable improvements.
- **Cepstral mean subtraction –** Primarily non-causal cepstral mean subtraction was performed over the entire train or test file. A few systems used causal ceptral mean subtraction (for example, in the form of RASTA filtering).

A few different and new approaches related to front-end processing were used in eval98 systems. These included (with the site using them in square brackets):

- **Nonlinear discriminant analysis (NLDA) features** [SRI] [13] – Here features are derived from using the hidden node outputs of a neural network trained to separate target and non-target cepstral features

- **Pitch prosodics** [SRI] – In addition to using the pitch frequency derived from each short-time window as an additional feature for the speaker model, the approach was aimed at modeling both the static and dynamic pitch contour information.
- **Modulation spectral filtering** [OGI] [26] – This approach looked at filtering the time sequence output from each spectral filterbank output to suppress the speaker-independent information in the spectral features .
- **Speaker-independent features and speaker mapping** [OGI] [11] – This approach is a tight integration between the feature extraction and the speaker modeling components. In general, a form of principal component analysis was used to derive features conveying speaker-independent, linguistic information from cepstral features and both were used to train a speaker-specific mapping function to characterize the speaker-dependent portion of the feature space.

Overall, little new work on features was deployed in eval98 systems. The NLDA features did provide a small but consistent gain over baseline cepstral features and pitch prosodics. While not powerful by themselves, NLDA features did boost performance when combined with cepstral feature scores. The other new features showed no improvement over standard cepstra.

4.3. Speaker Models

The approaches to speaker modeling used by various eval98 systems can be broadly categorized into two approaches based on the representation detail employed. The first category is termed unsupervised acoustic representation in which a speaker's acoustic characteristics (as reflected in the sequence of feature vectors extracted from his/her speech signal) are modeled by a single model with only implicit representation of underlying acoustic classes (such as broad-category sounds like vowels or fricatives). This is the case, for example, when a speaker representation is a model of the probability distribution of his/her features; the model represents the agglomeration of all speaker-specific sounds with no explicit modeling of individual sounds or sound classes.

The second category is termed supervised acoustic representation in which explicit segmentation, labeling and representation of underlying acoustic classes are used to model a speaker's acoustic characteristics. A

system of this type would use an external labeler, such as a phone recognizer, to segment and label a speaker's training speech so that speaker-specific models of each acoustic class (phoneme) could be trained. During recognition, the same external labeler would segment and label test speech so labeled test speech segments could be compared to speaker-specific label models.

In addition to the speaker model representation, how the background model is constructed and how the target speaker model and background model scores are combined to produce the likelihood ratio statistic are key to system performance. In the following two subsections we outline the specific modeling approaches for both types of representations, detailing the construction of the background model and the likelihood ratio statistic. As above we note the specific site using the approach in square brackets.

4.3.1. Unsupervised Acoustic Representations

As expected for a text-independent task, unsupervised acoustic models were the predominant models used in systems. Of these, Gaussian mixture models, especially adapted GMMs, were the models most often used primarily due to their modest computational requirements and consistently high performance.

- **Adapted Gaussian Mixture Models** [MITLL, Dragon, OGI, SRI] [24] – For this approach, the background model is a single, speaker-independent Gaussian mixture model (GMM) with 1024-2048 mixtures and a target speaker model is derived from the background model using Bayesian adaptation. The likelihood ratio statistic is then simply the ratio of the target to background models likelihood scores for a test utterance. This model has been the most widely used by sites over the past few NIST evaluations.

- **Unadapted Gaussian Mixture Models** [BBN, CIRC, ENST, IRISA] [22] – In this approach, each speaker is represented by independent GMMs, typically 32-128 mixtures, derived via maximum likelihood estimation from their training data. The background model is either a single, speaker – independent GMM or a collection of speaker-dependent GMMs termed cohorts, likelihood-ratio sets, or background sets. When using background sets, the background score for a test utterance is typically the arithmetic or geometric average of each set member's likelihood score. The likelihood ratio statistic is the ratio of the target to background likelihood scores for a test utterance.

- **Ergodic Hidden Markov Models** [A2RT] [12] – For this approach, the background model is a single, speaker-independent 4 state, 32 mixture/state, ergodic HMM and a target speaker model is derived from the background model using adaptation. The main difference of this model and the adapted GMM is the ergodic HMM explicitly models transition probabilities between hidden states. The likelihood ratio statistic is then simply the ratio of the target to background models likelihood scores for a test utterance.

- **Unimodal Gaussian Models** [LIA] [1] – In this approach, each speaker is represented by an independent, unimodal, full covariance Gaussian model with the background model being a speaker-independent unimodal, full-covariance Gaussian as well. For the implementation used in the eval98 system this simple model was augmented by more complicated processing such as using a sequence of fixed-length segments from each test utterance and a parallel bank of multi-band recognizers. The segment, band likelihood ratio was simply the ratio of target to background likelihood scores; the final test utterance score was a more complicated merging of individual likelihood ratio scores.

- **Auto Regressive Vector Models** [LIP6] [16] – In this approach, each speaker is represented by an independent ARVM, which is a predictive model of the sequence of spectral feature vectors. A collection of ARVMs from nontarget speakers is used as the background model. The likelihood ratio score is the difference between the target and best scoring background ARVM score on a test utterance.

4.3.2. Supervised Acoustic Representations

The use of supervised acoustic models have made a strong showing in this and previous evaluations. The major limitation for text-independent applications is in the accuracy and consistency of the labeler. Under more controlled conditions of text-dependent or vocabulary constrained applications, such systems are indeed the approach of choice. With the addition of better labelers and new scoring techniques, these approaches may outperform the unsupervised approaches for text-independent tasks.

- **Large Vocabulary Continuous Speech Recognizer** [Dragon] [8] – In this approach, the labeler used is a speaker-independent LVCSR system segmenting and labeling phoneme units. For each label, a

speaker-independent, monophone unit (3 state/128 mixtures) is used as a background model and target label models are derived from these background label models using Bayesian adaptation. Likelihood ratios for each label are computed as the ratio of target to background label model likelihood scores and the sequence of label likelihood ratios is integrated over the entire test utterance. In addition to using standard HMM-based techniques to compare train and test label segments, Dragon also examined a model free, sequential, non-parametric (SNP) technique to compare label segments from training and testing speech. When combined with their adapted GMM system scores, the SNP approach showed a significant improvement in performance.

- **Broad Phonetic Class Recognizer** [BBN, Ensigma] [3] – In the Ensigma system, a speaker-independent, ergodic, broad-class phonetic recognizer with 28 phonetic classes is used for segmentation and labeling. Background label models are 3 state/3 mixture HMMs and target label models are derived from the background label model via adaptation. The likelihood ratio score is computed not from the acoustic likelihood scores, but from the ratio of number of target label matches to number of background label matches from a Viterbi decoding of the test utterance. In addition to the standard technique of only scoring test speech against models derived from training speech, Ensigma examined a symmetric scoring technique that also scored training speech against models derived from test speech with some significant improvement. In the BBN system, a full speech recognition system was used to transcribe the speech into words and then the words were expanded into 53 phonemes using a dictionary expansion. The output phone sequence was then used as the input to a secondary discrete HMM to model the speaker dependent characteristics of the phoneme sequence. By itself this approach was not very effective and even when combined with GMM acoustic scores it showed no improvement in performance..

- **Temporal-Decomposition with Neural Nets** [CIRC] [10] – In this approach, the labeler is a blind, temporal-decomposition segmentation followed by a vector quantization labeling of segments into 8 label classes. Unlike the LVCSR or phonetic class approaches, the labels in this approach are acoustically rather than linguistically defined units. Label models are 3 layer multi-layer-perceptrons (5 frame input, 20 hidden nodes, and 2 output nodes) trained to discriminate between label-class features

from target and nontarget speakers. This system represents the only pure discriminant classifier used in the evaluation (SRI fielded an adapted GMM system that used secondary discriminant training). The test utterance score is the average of the MLP outputs over all observed label segments in the utterance.

4.4. Score Normalization/Microphone Compensation

One of the major challenges for the NIST speaker recognition evaluations is dealing with the channel variability found in telephone speech, primarily variability and distortions imposed by different microphones. While standard channel compensation techniques applied to spectral features, such as cepstral mean subtraction, do help, there is still a significant gap between performance under matched and mismatched microphone conditions. Three general approaches have emerged from the NIST evaluations to address this problem.

4.4.1. Znorm/Hnorm

For the NIST evaluation, systems were required to produce scores for which speaker-independent thresholds could be applied. Although likelihood ratio scoring does provide relatively stable speaker independent scores, there still exist biases from the data and models that imparts speaker-dependency on model scores. In the *znorm* compensation approach, these speaker-dependent score biases and spreads are estimated by observing the distribution of scores produced from speaker models scoring nontarget development speech. During testing a model score is adjusted using its estimated bias and spread parameters. It has been further observed that the microphone type (carbon-button or electret) used during training a model can also induce strong biases on scores. For example, a model trained using speech from an electret microphone will tend to produce higher likelihood ratio scores for speech from electret microphones regardless of the speaker. In an extension to znorm, a technique called *hnorm* was developed which estimated speaker and handset-dependent bias and spread parameters for compensation. The application of hnorm requires that putative handset labels be available for development, training and testing data. For eval98, NIST supplied handset labels to participants for all data. A majority of eval98 systems used either *znorm* or *hnorm* score normalization [23, 24].

4.4.2. Handset type mapping

To compensate for handset differences in the speech signal domain, MITLL [21] developed and applied a non-linear mapper for mapping electret speech into carbon-button speech and vise-versa. The mapper was applied during training to synthetically augment single-handset training data and during testing to map test speech from one microphone into speech from the training microphone. While showing improvement over a baseline system, this approach did not work as well as *hnorm* on the eval98 test set.

4.4.3. Handset type consideration in background models

Several sites used the provided handset labels to select the data for background model training or the members of cohort sets [9]. For a single speaker-independent background model a balance of electret and carbon-button speech was used to derive a microphone-independent model or separate microphone-dependent models were trained and associated with target models of the same microphone type. For cohort sets, microphone balanced speakers were selected and microphone dependent sets used for target speakers.

4.5. Fusion Systems

In this evaluation, several sites showed some improvement over a baseline system by simple linear combination of scores from different systems. For example, Dragon showed substantial performance improvement by combining scores from their adapted GMM system and their SNP system; SRI found additional gain in baseline performance by combining scores from their systems using cepstral, NLDA and prosodic features; Ensigma showed performance improvement by combining forward and backward scores in their symmetric scoring system.

Both the ELISA consortium and NIST performed true cross-site system fusion. IDIAP and ENST collaborated in using a Dempster-Shafer based fusion system to combine scores from consortium systems. The combined system, unfortunately, showed no improvement over the best consortium system. NIST examined two fusion systems combining scores from all eval98 participants. The first NIST fusion system was a simple voting system which looked at the hard decision outputs from each system an produced a score which was the average number of "true" decisions. This fused system did show improved performance over the best individual system in the low false-alarm region of the DET curve. The second

NIST fusion system utilized an MLP (multi-layer perceptron). Since no development data was available to find optimal MLP parameter settings, a jackknife experiment was conducted for robust parameter selection while testing on the entire test set. This fused system too had mixed results compared to the best individual system, doing better in some operating regions and worse in others (section 5.1 and also [7]).

5. The 1998 Evaluation: Results

The twelve research sites participating in the 1998 evaluation submitted results for a total of twenty-seven systems. NIST produced scoring results consisting of DET Curves for the specified training and test conditions, and for portions of the test data corresponding to various factors of interest. Some of these are described below. These and some additional results were presented at the RLA2C Workshop [19] and are available on the NIST website [18].

NIST analyzed various conditions that could be affecting recognition performance, including sex, age, pitch, handset type, noise, numbers of calls made, dialect region, and channel. For each of these conditions, we partitioned the training, target test, and/or non-target test data based on condition-relevant values, and analyzed the results.

It has been NIST policy in the speaker recognition evaluations not to publicly rank the performance results of the different participating sites. It is for this reason that the DET curves presented in this section do not identify the systems being considered. Specific information about the performance of individual systems may be available from researchers at the participating sites.

5.1. Overall Performance

Figure 2 shows the DET Curve for one of the test conditions, namely that of two-session training, 30-second durations, with same number tests. (In general, we concentrate on the two-session training, 30-second duration results in what follows.) In addition to results for twelve primary systems (system identities have been suppressed), it also includes two NIST created fusion systems.

The NIST12 system is a simple voting system. We combine the hard decisions of twelve primary systems, assigning the segment score as the number of TRUES minus the number of FALSES. The hard decision is true if the segment score is greater than zero. This approach produces a modest benefit compared to the

other best systems, as may be seen in Figure 2. The NIST12 system is plotted as a set of discrete points, since it has only a limited number of operating points.

The NISTmlp system is an attempt to extend the ideas used in NIST's system [7] for combining results from multiple speech recognition systems. It uses the MIT Lincoln Laboratory LINKnet software to create a multi-layer perceptron that combines the likelihood scores from the twelve primary systems. As may be seen in both the bar chart and DET Curve of Figure 2, the NISTmlp system outperforms the other systems represented.

5.2. Training Conditions and Duration

The remaining figures in this and the next three subsections show DET plots of performance restricted to certain conditions for one of the systems in the 1998 evaluation. For all of these plots the one system was chosen to be broadly representative of the results for all systems included in the evaluation. It should be understood, however, that in each case there is considerable variation in performance results across systems. The general trends that are noted are stronger for some systems than for others, but are not strongly reversed in the performance of any of the systems.

Figure 3 shows the variation in performance by training condition for one system. The performance differences are in the expected direction. Using training data from two different conversations improves performance over using the same amount of data from only one conversation. With these same two conversations, but using all available data, thus increasing the total amount of training data by an average factor of 2.7, there is a rather smaller improvement. There is some variation across systems, with at least one showing no gain from the increased amount of training data. Thus multiple session training, if the human factors of the application permit it, may significantly benefit performance.

Figure 4 shows the variation in performance by duration for one system, with longer durations producing better performance, as expected. This effect was more pronounced in some systems, and less in others, but was always present. There is more gain going from 3 to 10 seconds than from 10 to 30, suggesting that the benefits from increased duration are limited beyond a certain point.

5.3. Handset Effects

Figure 5 shows the performance variation for one system on same number and different number tests. Since specific handset information was not available, same or different number is assumed to correspond to same or different handset, though there are undoubtedly exceptions. Note the very large performance difference for same and different numbers. A large difference occurs for all systems.

Analysis following the 1997 evaluation showed that one key source of difficulty for systems is the use of telephone handsets of different type (carbon-button or electret microphone). This was made possible by one participant, MIT-Lincoln Laboratory, making available to NIST its software that attempts to classify speech segments as coming from either carbon or electret handsets. In 1998 systems were given this handset classification information along with the test data. They could use this information as they wished, aware that it could be less than totally accurate. The results presented here, however, assume the correctness of this MIT classifier.

Figure 6 shows the 1998 DET Curves for the system used in Figure 5 with the different number tests broken down according to whether the training handset microphone type (electret or carbon-button) of the speaker (true or otherwise) being tested matches that of the test segment. Note the large gap in performance between the matched and mismatched cases. Again, this is a performance variation found in all systems in the evaluation.

Figure 7 breaks apart the two curves of Figure 6 according to the specific handset types involved. Thus for the matched case it shows results for electret training and test and for carbon training and test. For the mismatched case, it shows results for carbon training and electret test, and for electret training and carbon test. The numbers of tests included, both target (true speaker) and non-target (non-true speaker), are indicated. The major point to note here is that performance is better when the test segments are electret than when they are carbon. This held true for all systems.

5.4. Sex and Pitch

It has been NIST policy in recent evaluations, as discussed in Section 2, not to include any cross-sex tests, i.e. male hypothesized speakers when a female is actually talking or vice versa. Thus each test is really two

separate tests, one on male speakers and one on female speakers. It is natural to ask if there is a performance difference between the two.

The "all" lines in Figures 9 (for 1998) and 10 (for 1997) below suggest that the performance difference between the sexes is fairly small but with a trend toward better results for males. These figures show results for one particular site in each of these evaluations, but the same trend was observed for almost all systems.

We also examined in several ways the effect of speaker average pitch on performance. Since there were some differences, we here look at both 1997 and 1998 results. We obtained average pitch estimates using the Entropics [6] software "get_f0" and "pplain" functions.

The Switchboard-2 Corpus contains a larger percentage of high pitched (and young) speakers, especially among females in the Phase 1 (1997 data). Figure 8 shows the distributions of average pitch frequencies of the training data among the chosen female and male target speakers in 1997 (~400 total speakers), and 1998 (~500 total speakers).

We chose to look at specific high and low-pitched subsets of the speakers, for both males and females. We examined performance on same number tests when both targets and non-targets were restricted to the high or low 25% of average training pitch frequencies for each sex, as well as overall performance for each sex. (The vertical lines in Figure 8 show these 25^{th} and 75^{th} percentile pitch values for the distributions.) Figure 9 shows these results for 1998, and Figure 10 gives similar results for 1997, for the systems of one particular site.

It may first be noted that performance over all female or male speakers is not consistently better or worse than over the restricted pitch subsets. One might have imagined that restricting the pitch range would make the recognition problem more difficult, but this does not appear to be the case. Average pitch differences thus do not appear to be a major cue for distinguishing speakers.

There are performance differences, however, between high and low pitched speakers. In both years performance was better for low-pitched than high-pitched males. For females, however, the low-pitched speakers did better in 1997 while the high-pitched ones were better in 1998. Performance in these matters tended to be consistent for different systems while being inconsistent between years. Perhaps this has

something to do with the large number of high-pitched females in the 1997 test set, or perhaps these are just random effects. More investigation would be of interest.

We also examined partitioning the targets or the non-targets by "closeness" in pitch between the training data and the test segment. The more interesting results came from partitioning the targets in this way. Figure 11 shows these results for same number tests for one system in 1998.

The plot shows performance for all tests and for when target tests (true speaker tests) are restricted to the high or low 10% or 25% of all such tests based on pitch closeness. The results are all in the expected direction, but the big difference of note is the considerable degradation in performance when tests are restricted to when the model and test pitch values are far apart. For example, for a fixed miss rate in the 2-20% range, the false alarm rate generally doubles between the "all" case and the "25% far" case for the system used in Figure 11. This presumably corresponds to situations when, because of a cold or stress or other factors, a speaker does not talk quite as he or she normally does. While varying in magnitude, such a degradation occurs for all systems.

5.5. Other Factors

Noise was, not too surprisingly, found to also be a factor that moderately affected speaker recognition performance. Speakers were encouraged to initiate calls from different phones, resulting in a fair number from pay phones in outdoor or other noisy locations. The ten second segments were manually labeled as "good", "bad" or "really bad", and true speaker test scores were conditioned on this label. (Impostor scores were held the same for all of the conditions.) The results, shown in figure 12, demonstrate a modest correlation, with about a factor of two increase in error rate for "really bad" data over "good" data. Further study is needed to categorize which types of noise may most impact performance

Corpus speakers provided their age as part of the enrollment process. Most were of college age (late teens or early twenties), but there were some older speakers. As might be expected, false alarm rates were lower when the target and model speakers were further apart in age, but this effect was quite modest. Factors not found to have a significant effect on performance included how many calls a speaker made, place of birth (perhaps reflecting dialect) and conversation side; i.e., whether the speaker initiated or received the call.

6. Summary and Perspective

NIST has supported regular speaker recognition evaluations, open to all, with announced schedules, written evaluation plans, and follow-up workshops. These can be an effective means to encourage research and develop state-of-the-art systems in core technology areas such as speaker recognition. Furthermore, appropriate evaluation methodologies and analysis of results can help illuminate the progress that has been made and identify the factors limiting further advance. NIST intends to use the results of such analysis in designing future evaluations, where these and other factors will be studied further.

Any site or research group desiring to participate in future evaluations should contact Alvin Martin at NIST (alvin.martin@nist.gov) and should obtain data from previous evaluations from the LDC to develop and test their systems for evaluation tasks.

6.1. Future Research Directions

From an examination in Section 4 of the technologies used in the 1998 NIST evaluation, there are some observations we can make about the current status of speaker recognition research and possible future research directions.

In general, new features have provided small, but consistent gains in performance. The main goal is to find features that are more immune to the variability and degradations we know affect speaker recognition, such as channel, microphones and acoustic environment. It is highly unlikely that these new features will be derived from the spectrum since the spectrum is obviously highly affected by the above factors. Non-spectral features, such as pitch, have shown some promise of variability immunity, but are not very robust to other factors like emotions and in general are much less effective in speaker separation.

To date, supervised acoustic representations have not outperformed unsupervised acoustic representations for text-independent applications. These techniques generally require complicated labelers with large computational demands with limited returns in performance. Gains are likely to remain limited as long as the labeler is only used as an elaborate means to compute the acoustic likelihood of the data. The real payoff in these approaches is likely to be in using the label sequence output to learn about higher-levels of information not currently found in and complimentary to the acoustic score. Exploitation of such high-level

information may require some form of event-based scoring techniques, since higher-levels of information, such as indicative word usage, will not likely occur regularly as acoustic information does.

One of the largest robustness challenges for a speaker recognition system is dealing with mismatched conditions, especially microphone mismatches. Since most system rely primarily on acoustic features, such as spectra, they are too dependent on channel information. While this can be a plus for applications where speaker and microphone are very likely to be linked (e.g., facility access control), it is a large impediment to more general application deployment and a fundamental research challenge. It is likely that decoupling of the speaker and channel will come from a better understanding of specific channel effects on the speech signal since this would lead to the immediate payoff of better features and compensation techniques.

Fusion of systems may be a means to build on a solid baseline approach and provide the best attributes of different systems. This of course requires the search for complementary information to model; what indications are strong when the baseline is weak? Furthermore, successful fusion will require ways to adjudicate between conflicting signals and to combine systems producing continuous scores with systems producing event-based scores.

6.2. A Technology Perspective

Speaker recognition technology has made tremendous strides forward since the initial work in the field some 30 years ago. The perspective then was that, while the research was interesting, the technology would never be practical – because it would take a whole computer to perform the task! And of course this was back when computers cost real money, orders of magnitude more money than now, for a computer that by today's standards would be totally inept. If researchers then could possibly have imagined, and believed in, today's computing and information infrastructure, what would they have thought? Probably that making a business success of speaker recognition would be trivial! But since we're not there yet, it might be good to reflect on why not.

One of the problems that must be faced in the application of speaker recognition for access control is the rejection of valid users. In real systems, it is utterly unacceptable to reject valid users. Therefore there must be acceptable backup procedures to ensure that valid users are not rejected.

Another stumbling block for automatic speaker recognition is the discrepancy between human and machine performance. Historically, humans have outperformed machines by a wide margin. However, humans and machines give different results and seem to operate quite differently. For example, humans are more robust than machines to distortions and noise. On the other hand, machines have demonstrated greater ability than humans to distinguish the voices of identical twins. Perhaps we are approaching the point where machines overtake humans in speaker recognition ability.

Figure 13 shows the results of a recent comparison of human versus machine performance on three second test segments, made during the 1998 NIST evaluation of text-independent speaker recognition. While neither humans nor machines did an impressive job, the difference was not great – about a factor of 2 in error rate. (Note that both same number and different number tests are included in Figure 13. This is why this figure shows poorer machine performance than that shown in Figure 4. For a further discussion of human vs. machine performance on a limited subset of the full evaluation test set, see [25].)

Future directions in speaker recognition technology are not totally clear, but several observations might be helpful. First, computer power will continue to grow exponentially for at least the near (and foreseeable) future. The challenge is clear – figure out how to exploit that power, because it's a safe bet that major advances will demand ever greater computing power and that leading researchers will have figured out how to use it productively. Second, human listeners have a relatively keen ability to recognize familiar voices. (People apparently are much more accurate in classifying familiar voices than unfamiliar voices.) It might therefore be worthwhile to try to understand the nature of this capability and to begin to create models to perform this function by computer. This sounds like a formidable challenge. But at least we have an existence proof to encourage us onward!

NIST Speaker Recognition Evaluation Training and Test Conditions			
	Condition	**Description**	**Notes**
Training Data consists of concatenated speech segments derived from energy based speech detector	One-session	2 minutes of speech from one conversation	1997 and 1998
	Two-session	1 minute of speech from each of two same-number conversations	1997 and 1998
	Two-session full	all speech data from above two conversations	1998 only
	Two-handset	1 minute of speech from each of two different number conversations	1997 only
Test Data consists of concatenated speech segments derived from energy based speech detector	Sex	male and female	no cross-sex tests
	Duration (approximate)	3, 10, and 30 second segments	3 sec. is subsegment of 10 sec. which is subsegment of 30 sec.
	Same/different phone number	true speaker training vs. test segment	phone number serves as proxy for handset
	Same/different handset type	model speaker training vs. test segment	based on MIT handset labeler

Table 1: Training and test conditions in NIST 1997 and 1998 Speaker Recognition Evaluations

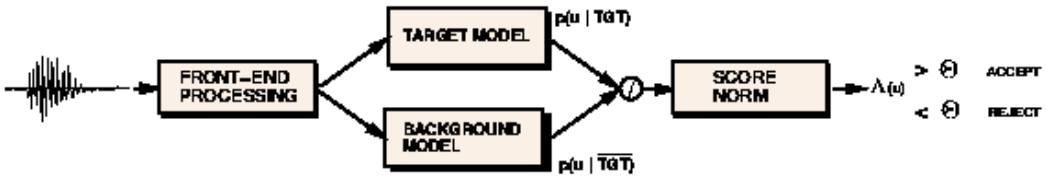

Figure 1: Canonical structure of a speaker recognition system. The general approach used by all systems in the NIST eval98 was that of a likelihood ratio comparison detector consisting of three main components: front-end processing, target and background modeling, and score normalization.

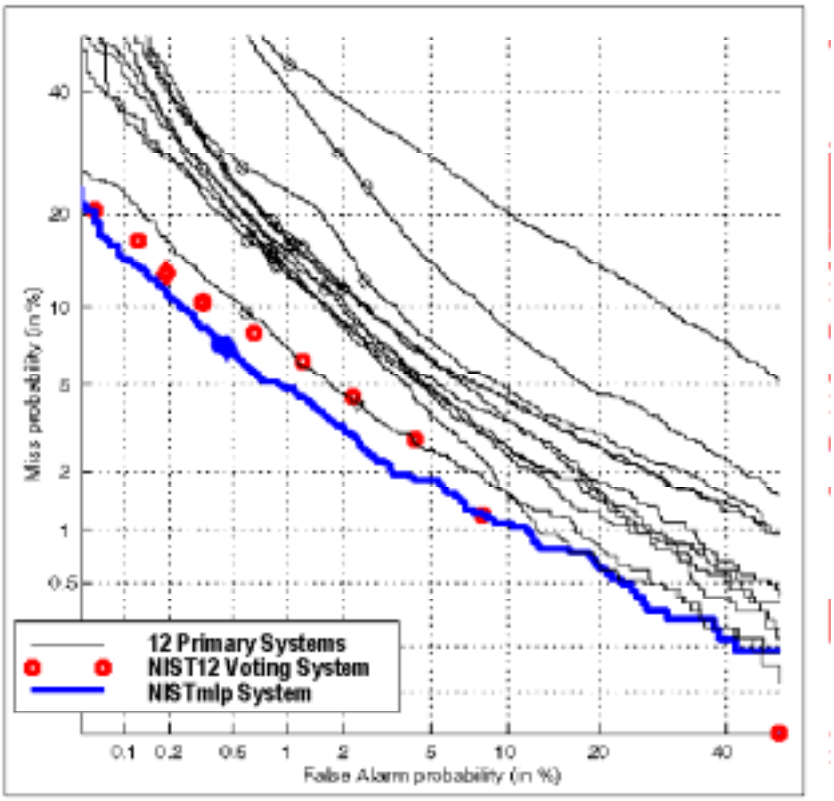

Figure 2: DET Curves for the primary systems of the twelve participants and two NIST fusion systems, processing the 1998 NIST Speaker Recognition evaluation data. Results shown are from same number, 30-second test segments using the two-session training condition.

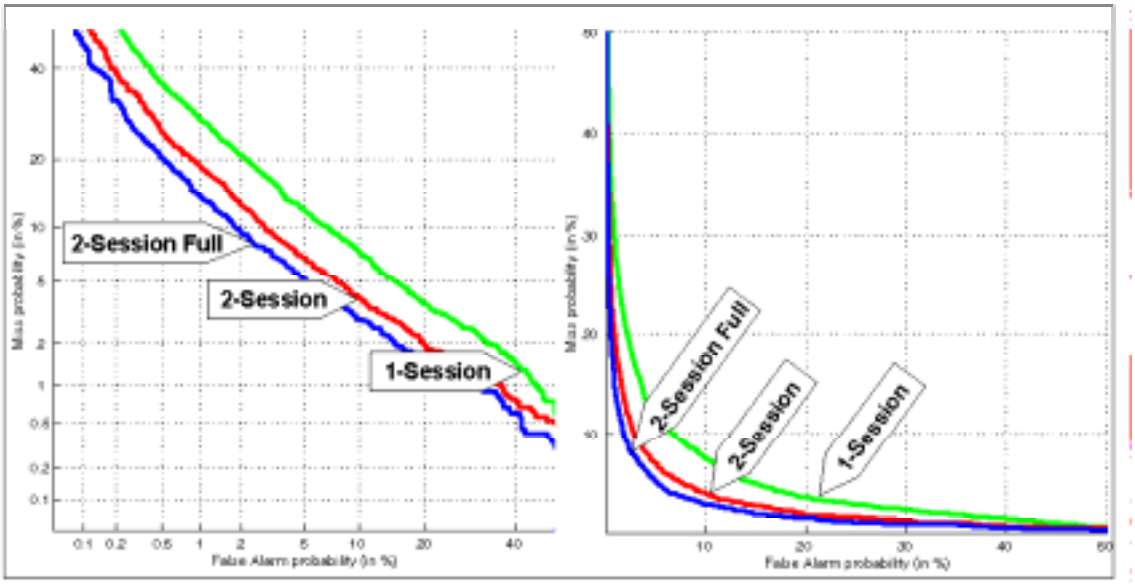

Figure 3: System performance is shown for each training condition processing the same test data (same number tests of 30-second durations). The DET plot is on the left. On the right for contrast is a traditional ROC plot of the same data. It is seen that multiple training sessions using the same amount of data improves performance. Additional training data improves performance slightly.

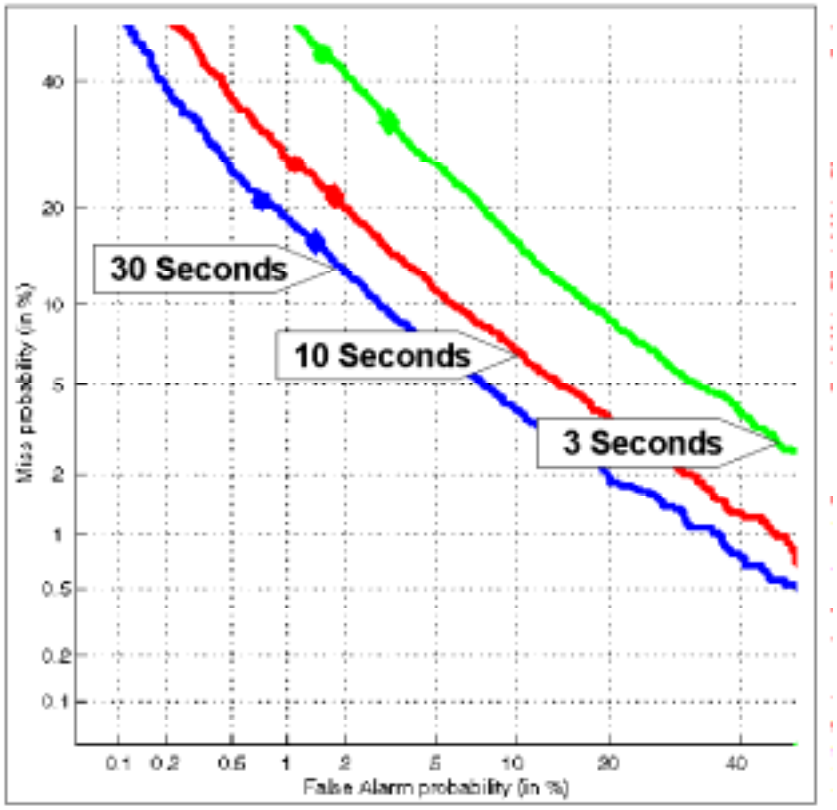

Figure 4: System performance is shown as a function of test segment duration under the same training condition (two-session). It is seen that the longer the test segment, the better the performance, however, with diminishing returns.

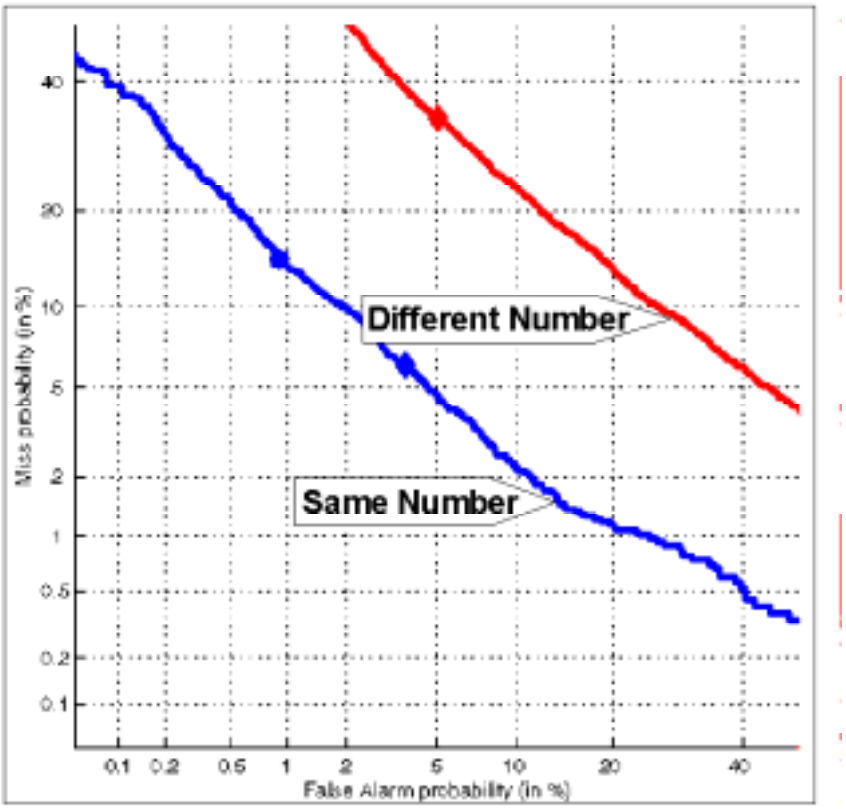

Figure 5: System performance is shown for same and different phone number test segments for the 2-session training condition, 30-second test segments. It is seen that phone number (and presumably handset) matching between training and test greatly improves performance.

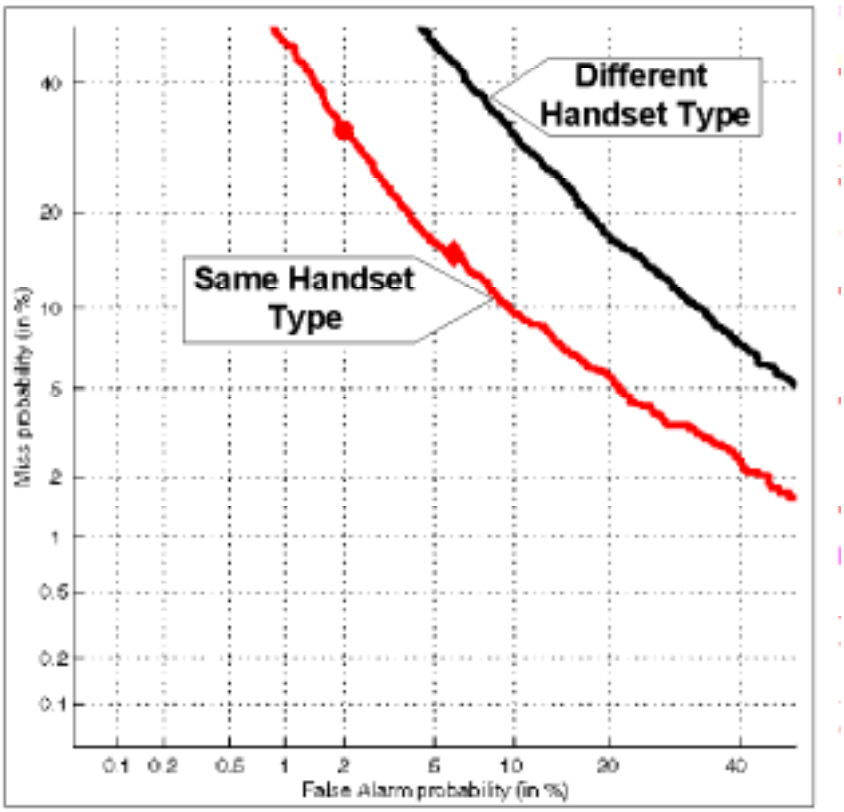

Figure 6: System performance is shown for the 2-session training condition, 30-second test segments when the test segment comes from a phone number that was not used for training. It is seen that handset type matching between training and test greatly improves performance.

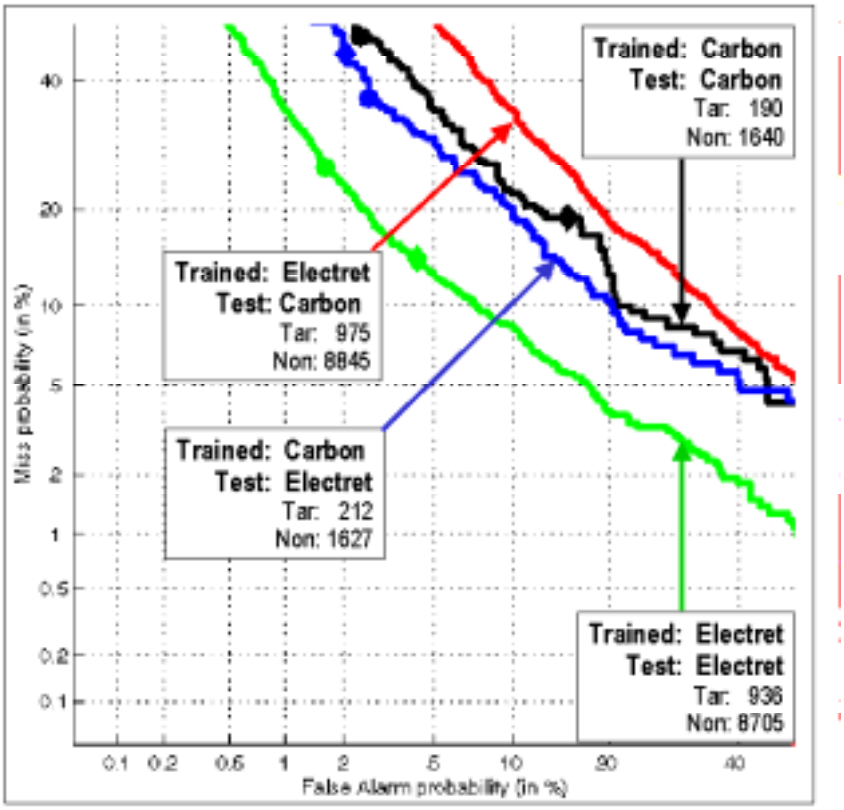

Figure 7: System performance is shown as a function of handset type for both matched and mismatched tests. The test was from the 1998 NIST Speaker Recognition evaluation using the two-session training condition with different number 30-second test segments. The general trend is that performance is best with electret test segments.

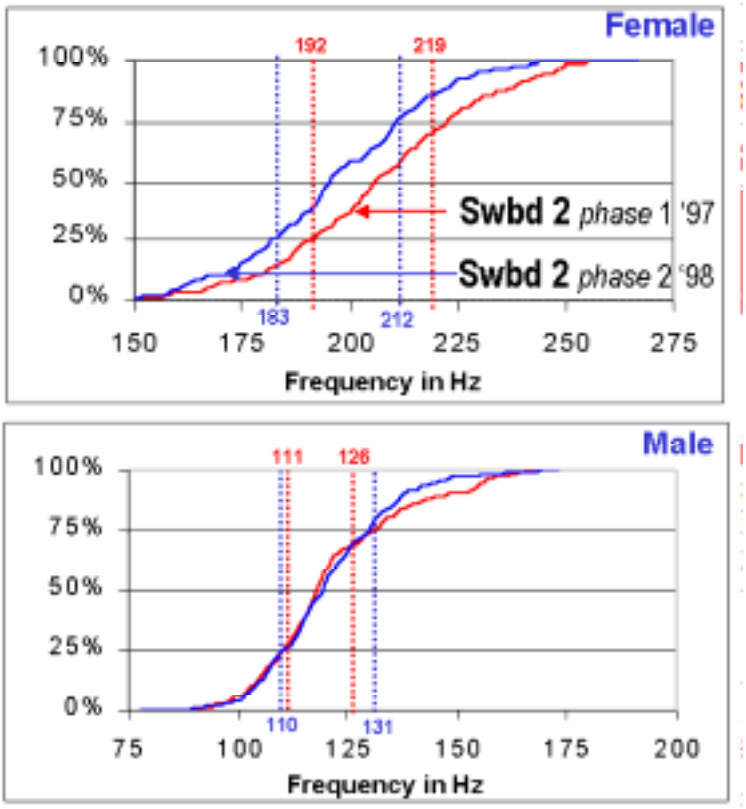

Figure 8: Distributions of average pitch over training data for female and male target speakers in 1997 (Switchboard-2 Phase 1) and 1998 (Switchboard-2 Phase 2) evaluations. The 25th and 75th percentile pitch values of each distribution are shown by the vertical lines. The test set for 1997 contained an unusually large number of high pitch females. There is not much disparity in the male pitch distribution from year to year.

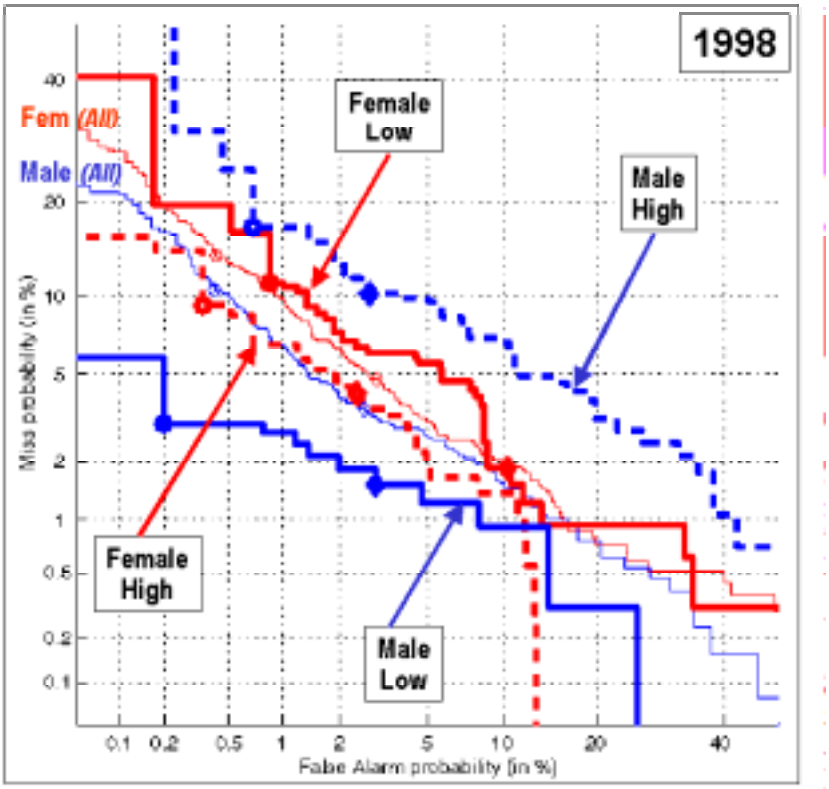

Figure 9: System performance is shown as a function of sex and for high and low pitched speakers. The upper and lower 25% distributions (shown in figure 7) are plotted for the two-session training condition and the same number 30-second test segments.

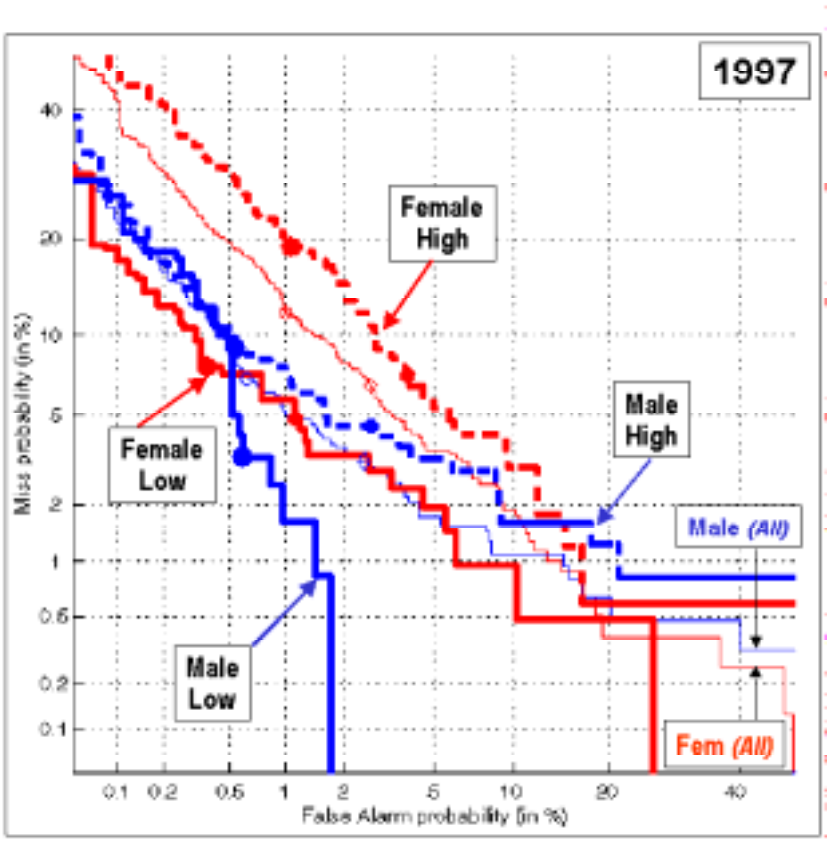

Figure 10: System performance is shown as a function of sex and for high and low pitched speakers. The upper and lower 25% of distributions (shown in figure 7) are plotted for the two-session training condition and the same number 30-second test segments.

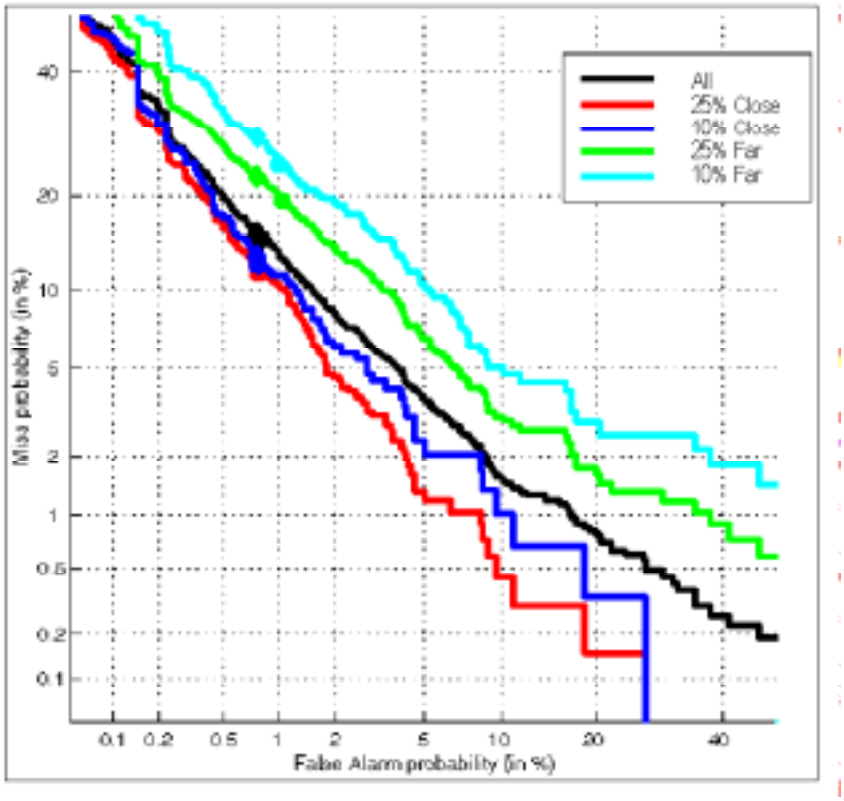

Figure 11: System performance is shown as a function of pitch closeness between the true speaker training data and the test segments. The results are shown for the two-session training condition on same number 30-second test segments. The high and low 10 percent and 25 percent in pitch closeness of all target tests are plotted, along with the curve for all such tests. It is shown that performance degrades when target models and test segments are not close in pitch.

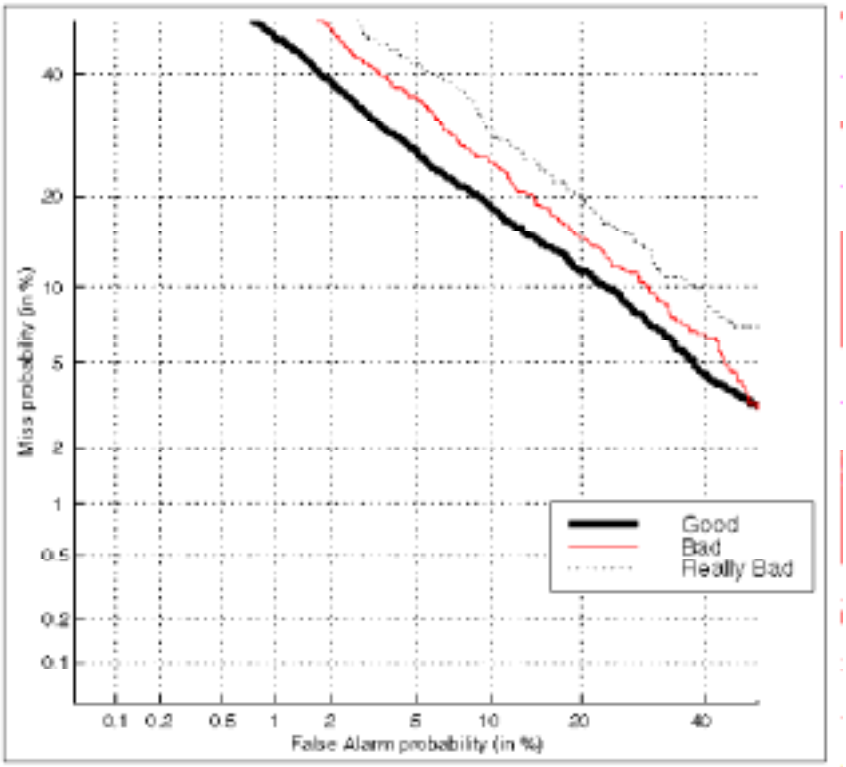

Figure 12: A DET plot for text independent speaker detection, contrasting true-speaker performance for tests segments subjectively classified as ""good", "bad" and "really bad".

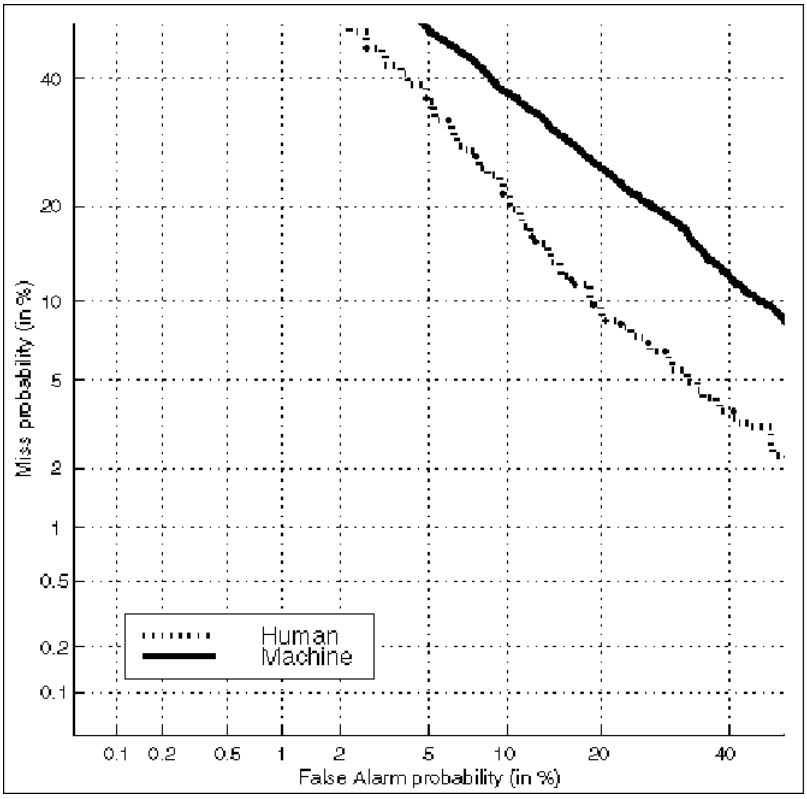

Figure 13: A DET plot for text independent speaker detection, contrasting human listener performance with machine performance for 3 second test segments. Both same number and different number tests are included in the curves shown.

References

[1] L. Besacier and J.-F. Bonastre, "Time and Frequency Pruning for Speaker Identification," RLA2C pp 106-110 April 1998

[2] J. Campbell, "Speaker Recognition: A Tutorial", Proc. IEEE, vol. 85, no. 9, Sept. 1997, pp. 1437-1462.

[3] M. Carey and E. Parris, "Cross Validation in Speaker Recognition," RLA2C pp 161-164 April 1998

[4] G. Doddington, "Speaker recognition evaluation methodology: a review and perspective", Proc. RLA2C, Avignon, 20-23 April 1998, pp. 60-66.

[5] G. Doddington et al., "Sheep, Goats, Lambs, and Wolves: A Statistical Analysis of Speaker Performance in the NIST 1998 Speaker Recognition Evaluation", Proc. ICSLP '98.

[6] Entropic Research Laboratories, 600 Pennsylvania Ave. SE Suite 202 Washington DC 20003.

[7] J. Fiscus, "Recognizer Voting Error Reduction", Proc. *IEEE Workshop on Automatic Speech Recognition and Understanding,* Dec. 1997.

[8] L. Gillick et. al, "Application of Large Vocabulary Continuous Speech Recognition to Topic and Speaker Identification using Telephone Speech," ICASSP pp II-471-II-474, April 1993

[9] L. Heck and M. Weintraub, "Handset-dependent Background Models for Robust Text-Independent Speaker Recognition," ICASSP pp. 1071-1073 April; 1997

[10] J. Hennebert and D. Petrovska-Delacretaz, "Phoneme-Based Text-Prompted Speaker Verification with Multi-layer Perceptrons," RLA2C pp 55-58 April 1998

[11] H. Hermansky and N. Malayath, "Speaker Verification using Speaker-Specific Mappings," RLA2C pp 111-114, April 1998

[12] C. Jaboulet, J, Koolwaaij, J. Lindberg, J.-B. Pierrot and F. Bimbot, "The CAVE-WP4 Generic Speaker Verification System," RLA2C pp 202-205 April 1998

[13] Y. Konig, L. Heck, M. Weintraub, and K. Sommez. "Non-linear Discriminant Feature Extraction for Robust Text-independent Speaker Recognition," RLA2C pp 72-75 April 1998

[14] Linguistic Data Consortium (LDC), Philadelphia, PA, USA, www.ldc.upenn.edu

[15] A. Martin et al., "The DET Curve in Assessment of Detection Task Performance", Proc. *EuroSpeech '97,* pp. 1895-1898, Sept. 1997.

[16] C. Montacie and J. Le Floch, "AR Vector Models for Free-Text Speaker Recognition," ICSLP pp 611-614, Banff, Canada 1992

[17] NIST 1997 Speaker Recognition Evaluation Plan, www.nist.gov/speech/sp_v1p1.htm.

[18] NIST 1998 Speaker Recognition Evaluation Plan, www.nist.gov/speech/spkrec98.htm.

[19] M. Przybocki and A. Martin, "NIST speaker recognition evaluation - 1997", Proc. RLA2C, Avignon, 20-23 April 1998, pp. 120-123.

[20] M. Przybocki and A. Martin, "NIST Speaker Recognition Evaluations", Proc. LREC, Granada, Spain, 28-30 May 1998, vol. 1, pp. 331-335.

[21] T. Quatieri, D. Reynolds, and G. O'Leary "Magnitude-Only Estimation of Handset Nonlinearity with Application to Speaker Recognition," ICASSP pp 745-748 May 1998

[22] D. Reynolds, "Speaker Identification and Verification using Gaussian Mixture Speaker Models," Speech Communication v17, pp 91-108 August 1995

[23] D. Reynolds, "HTIMIT and LLHDB: Speech Corpora for the Study of Handset Transducer Effects," ICASSP pp 1535-1538 April 1997

[24] D. Reynolds, "Comparison of Background Normalization Methods for Text-Independent Speaker Verification," Eurospeech, pp 963-967 September 1997

[25] A. Schmidt-Nielsen and T. Crystal, "Human vs. Machine Speaker Identification with Telephone Speech", Proc. ICSLP '98.

[26] S. van Vuuren and H. Hermansky, "!MESS: A Modular, Efficient Speaker Verification System," RLA2C pp 198-201, April 1998